

Statistik i basketball

En note til opgaveskrivning*

Jerôme Baltzersen
jerome@falconbasket.dk

14. marts 2010

Indledning I Falcon og andre klubber er der en del gymnasieelever, der på et tidspunkt i løbet af deres gymnasietid skal skrive en større opgave. Er man interesseret i matematik, og ønsker man at kombinere det med ens interesse for basketball, er en opgave om anvendelser af statistik i basketball en spændende mulighed. Der findes mange udmærkede bøger om statistik på alle niveauer, men der er mig bekendt ingen litteratur på dansk, der behandler konkrete anvendelser på scoringsprocenter i basketball. Dette tomrum forsøger denne lille note at udfylde. Den er altså rettet mod gymnasieelever, og bør ses som et supplement til de gymnasiale lærebøger. Sportsinteresserede i almindelighed med en interesse for statistik vil dog givetvis også kunne have glæde af noten. Et matematisk niveau svarende til gymnasiet samt lysten til at lære stoffet er eneste forudsætning.

Binomialfordelingen En spillers scoringsprocent kan angives som et tal $p \in [0, 1]$. Antager vi at scoringsforsøgenes udfald ikke påvirker hinanden (at de er uafhængige) kan vi opfatte dette som at et forsøg på at score vil lykkes med sandsynlighed p . Man kan så spørge om, hvad sandsynligheden er for at en spiller med scoringsprocent p scorer på netop x ud af n skud eller formuleret matematisk: Hvis X er en *stokastisk variabel*, der angiver antallet af scoringer ud af n forsøg, hvad er da $P(X = x)$. *Binomialfordelingen* giver svaret

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad (1)$$

hvor $\binom{n}{x}$ kaldes *binomialkoefficienten n over x* og er givet ved

$$\binom{n}{x} = \frac{n!}{x!(n-x)!},$$

hvor $k! = 1 \cdot 2 \cdot \dots \cdot (k-1) \cdot k$.

EKSEMPEL Med en scoringsprocent er 50 % er sandsynligheden for at score på 3 ud af 5 skud $5/16$.

*Tak til Morten Hornbech for mange kommentarer samt hjælp til at gøre formuleringerne mere klare.

Statistisk model En statistisk model består af en mængde af mulige observationer (udfaldsrummet) og en mængde af mulige fordelinger for disse observationer. En fordeling tildeler hver observation en sandsynlig således at summen bliver 1. Ser vi på situationen fra før er mængden af mulige observationer givet ved $E = \{0, 1, \dots, n\}$, altså de mulige antal gange man kan score på n skud. Mængden af mulige fordelinger er givet ved binomialfordelingerne for alle $p \in [0, 1]$, da vi forestiller os at vi ikke på forhånd kender spillerens scoringsprocent. Den korte notation for modellen er

$$(E, (P_p)_{p \in [0,1]}), \quad (2)$$

I denne model kan man ud fra relevant skudstatistik undersøge om en given spiller eller et givent hold kan antages at have scoringsandsynlighed p .

EKSEMPEL Simon har skudt 25 straffekast og scoret på 17 (vores observation). Er det rimeligt at antage at Simons scoringsandsynlighed er 0,75?

Man kunne også være interesseret i at sammenligne to spillere/hold. I dette tilfælde bliver den statistiske model

$$(E, (P_{(p_1, p_2)})_{(p_1, p_2) \in [0,1]^2}), \quad (3)$$

med $E = \{0, \dots, n_1\} \times \{0, \dots, n_2\}$, hvor n_1 og n_2 er antallet af scoringsforsøg for den ene henholdsvis den anden spiller/hold, og

$$\begin{aligned} P_{(p_1, p_2)}(X_1 = x_1, X_2 = x_2) &= \binom{n_1}{x_1} p_1^{x_1} (1 - p_1)^{n_1 - x_1} \cdot \binom{n_2}{x_2} p_2^{x_2} (1 - p_2)^{n_2 - x_2} \\ &= \prod_{r=1}^2 \binom{n_r}{x_r} p_r^{x_r} (1 - p_r)^{n_r - x_r}. \end{aligned}$$

Det forudsættes her at alle observationerne er uafhængige.

EKSEMPEL Falcons damehold har i første halvdel af sæsonen skudt 139 3'ere og ramt på 57, mens de i anden halvdel af sæsonen har skudt 109 3'ere og ramt på 56. Er Falcon blevet bedre til at skyde i løbet af sæsonen?

Estimation Ofte er man interesseret i at estimere dvs. "gætte" en given størrelse – fx et holds skudprocent – så godt som muligt. Det bedste gæt (det mest sandsynlige valg) betegnes maksimaliseringsestimatorens, som betegnes med \hat{p} . I binomialmodellen er $\hat{p} = \frac{x}{n}$, hvilket intuitivt ikke overrasker (hvorfor ikke? Kan du bevise det matematisk?).

Test og hypoteser

Signifikansniveau Man vælger på forhånd ens *signifikansniveau* ofte betegnet α . Man vælger stort set altid $\alpha = 0.05$, uden at nogen ville påstå, at det er bedre end så meget andet.

Man kan tænke på signifikansniveau som følger: Man forestiller sig, at man kan gentage eksperimentet et utal af gange. Hvis den situation vi ser i vores data

forekommer mindre end 5 % af gangene vil vi sige, at det er for ekstremt til at være i overensstemmelse med vores model (under hypotesen) og vi forkaster den.

EKSEMPEL En straffekastsskytte skyder 10 skud og rammer på 1 skud. Vi vil undersøge hypotesen om, at hendes scoringsprocent er 70 %. Hypotesen ender med at blive forkastet af følgende grund: Forestiller vi os, at hun ville have en scoringsprocent på 70 % og vi 100 gange satte hende til at skyde 10 straffekast ville det under 5 % af gangene ske, at hun kun ville ramme et skud. Derfor tror vi ikke på, at hendes scoringsprocent er 70 % og hypotesen forkastes.

Hypoteser Typisk vil en hypotese være “der er forskel på hold A og hold B’s scoringsprocenter.” I statistik undersøger man dog altid hypoteser på formen “der er *ingen* forskel på hold A og hold B’s scoringsprocenter,” og ser så om man kan forkaste denne hypotese. Her gælder det om at holde tungen lige i munden: Man viser altså ikke, at der er forskel, men afviser i stedet på et givet signifikansniveau, at der ingen forskel er. Læs lige det igen!

En hypotese i modellen givet ved (2) kaldes en *simpel* hypotese og skrives som:

$$H_0 : p_1 = p. \quad (4)$$

Derimod kaldes en hypotese i modellen givet ved (3) for en *sammensat* hypotese:

$$H_1 : p_1 = p_2 = p. \quad (5)$$

Teststørrelse Der findes et utal af teststørrelser man kan bruge, hvor en ofte brugt er *kvotientteststørrelsen*, $Q(x)$. For en simpel hypotese har vi (4)

$$Q(x) = \left(\frac{\hat{p}}{p}\right)^x \cdot \left(\frac{1-\hat{p}}{1-p}\right)^{n-x} = \left(\frac{x/n}{p}\right)^x \cdot \left(\frac{1-x/n}{1-p}\right)^{n-x},$$

mens den for sammenligning af to grupper, dvs. sammensat hypotese (5), er

$$Q(x) = \prod_{r=1}^2 \left(\frac{x_{\bullet} n_r}{x_r n_{\bullet}}\right)^{x_r} \left(\frac{(n_{\bullet} - x_{\bullet}) n_r}{(n_r - x_r) n_{\bullet}}\right)^{n_r - x_r}, \quad (6)$$

hvor $n_{\bullet} = n_1 + n_2$ og $x_{\bullet} = x_1 + x_2$. En teststørrelse er en hjælpestørrelse, der skal hjælpe os med at sige noget om, hvor godt en observation passer med vores hypotese.

Testsandsynligheden Kaldes også for p -værdien for et test, og betegnes med $\epsilon(x)$ for en given observation x . Følgende er sprogligt kludret så læs det langsomt!

Testsandsynligheden udtrykker sandsynligheden for at få en observation, der passer dårligere eller lige så dårligt, som det vi har observeret under antagelsen om, at hypotesen H er opfyldt.

Lad os sige, at $\epsilon(x) = 0.5$. Det betyder altså, at hvis vi udfører eksperimentet et stort antal gange vil vi i gennemsnit 50 % af gangene få en observation, der passer dårligere. Således må vores konkrete observation x altså passe ganske godt. *Jo større testsandsynligheden er jo “mere rigtig” er ens hypotese.*

Når man på forhånd har fastlagt et signifikansniveau α , forkaster man hypotesen H på niveau α , hvis $\epsilon(x) \leq \alpha$.

Approksimation af testsandsynlighed Man kan ofte godt finde testsandsynligheden eksakt, men det er utrolig besværligt. Heldigvis gælder følgende tilnærmelse/approksimation – der meget hurtigt bliver ret god – for store n_1 og n_2 (i praksis bare større end 15-stykker):

$$\epsilon(x) \approx 1 - F_{\chi_1^2}(-2 \log Q(x)).$$

Her er $F_{\chi_1^2}$ fordelingsfunktionen for χ^2 -fordelingen¹ med 1 frihedsgrad, mens “log” betegner den naturlige logaritme (ofte “ln” i gymnasiet).

Man får selvfølgelig et tal ud af $-2 \log Q(x)$, der i øvrigt skal være positivt (ellers har man tastet/regnet forkert!). Dette sættes så ind i en (computer) tabel for χ^2 -fordelingen, som man fx kan finde på <http://www.medcalc.be/manual/chi-square-table.php>.

Pas på! Tabellen angiver $1 - F_{\chi_1^2}(y)$, altså testsandsynligheden, og i øvrigt står “DF” for frihedsgrader (degrees of freedom). Vi er på jagt efter det y hvor $1 - F_{\chi_1^2}(y) = 0.05$. Er nemlig y større må vi forkaste vores hypotese.

Ifølge tabellen er grænsen $y = 3.841$ for én frihedsgrad. Det vil sige, at hvis $-2 \log Q(x) > 3.841$ må vi forkaste hypotesen.

Gennemregnet eksempel

Hold A har skudt 738 og ramt 342. Hold B har skudt 584 og ramt 238. Den statistiske model bliver

$$(E, (P_{(p_1, p_2)})_{(p_1, p_2) \in [0, 1]^2})$$

med $E = \{0, \dots, 738\} \times \{0, \dots, 342\}$. Udfaldsrummet er altså alle tænkelige kombinationer af talpar, hvor 1. koordinaten ligger mellem 0 og 738, mens 2. koordinaten ligger mellem 0 og 584: fx (32, 341), men ikke (342, 603). Videre er

$$P_{(p_1, p_2)}(X_1 = 342, X_2 = 238) = \binom{738}{342} p_1^{342} (1 - p_1)^{738 - 342} \cdot \binom{584}{238} p_2^{238} (1 - p_2)^{584 - 238},$$

hvor $p_1 \in [0, 1]$ og $p_2 \in [0, 1]$. Vores hypotese er:

$$H_0 : \text{Der er ingen forskel på hold A og hold B skudprocent.}$$

Vi finder kvotientteststørrelsen ved at sætte ind i (6), og får $Q(x) = 0.126119$. Dermed bliver $-2 \log Q(x) = 4.1311$. Dette er klart større end 3.831, så hypotesen forkastes med et brag. Der er altså forskel!

Vores bedste estimat for scoringsprocenterne for hold A er $\hat{p}_1 = \frac{342}{738} \approx 46\%$, mens vi for hold B finder $\hat{p}_2 = \frac{238}{584} \approx 40\%$. At vi finder, at der er en forskel er betryggende, da der er 6 procentpoints forskel, og begge hold har skudt et stort antal skud.

Hvad mangler?

Det er selvfølgelig ikke muligt på så lidt plads at behandle alt; så hvad mangler? Der mangler motivationen for, hvorfor vi gør som vi gør. Det kan virke oplagt og

¹Udtales “ki-i-anden”. Hvis man sammenligner k grupper skal man bruge χ^2 -fordelingen med $k - 1$ frihedsgrader.

rigtigt, men der er også andre måder at gøre det på. Når man så har fastlagt sig på principperne skal man bevise, at det rent faktisk ender med at blive de formler vi har brugt. Hvis man har lyst til at fylde mere matematisk indhold i sin opgave er dette en mulig vej at gå.

Ønsker man derimod at diskutere baggrunden for statistik, er der også masser at hente. Spørgsmål som “hvilket signifikansniveau skal man bruge?”, “hvorfor overhovedet vælge signifikansniveau? Kan man ikke bare finde ens testsandsynlighed og så se om man er tilfreds?”, “hvad betyder det, at man forkaster en hypotese?”, “hvad er type I og type II fejl?” og mange flere danner et godt grundlag for en omfattende diskussion af statistik.

Et andet populært tema er de såkaldte *player efficiency ratings*, der i mange ligaer bruges til at vurdere, hvor produktive/effektive en spiller er. Dette emne er uddybende behandlet i *Matheletics* af Wayne Winston.

SPØRGSMÅL, RETTELSE OG KOMMENTARER modtages meget gerne via mail på jerome@falconbasket.dk.